UC Riverside

UC Riverside Previously Published Works

Title

RIPEx: Extracting Malicious IP Addresses from Security Forums Using Cross-Forum Learning

Permalink

https://escholarship.org/uc/item/5h18109p

Journal

CoRR, 10939

ISSN

0302-9743

Authors

Gharibshah, Joobin

Papalexakis, Evangelos E

Faloutsos, Michalis

Publication Date

2018

DOI

10.1007/978-3-319-93040-4_41

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

arXiv:1804.04760v1 [cs.IR] 13 Apr 2018

RIPEx: Extracting malicious IP addresses from security forums using cross-forum learning

Joobin Gharibshah, Evangelos E. Papalexakis, and Michalis Faloutsos

University of California - Riverside, CA
900 University Ave, Riverside, California 92557
jghar002,epapalex,michalis@cs.ucr.edu

Abstract. Is it possible to extract malicious IP addresses reported in security forums in an automatic way? This is the question at the heart of our work. We focus on security forums, where security professionals and hackers share knowledge and information, and often report misbehaving IP addresses. So far, there have only been a few efforts to extract information from such security forums. We propose RIPEx, a systematic approach to identify and label IP addresses in security forums by utilizing a cross-forum learning method. In more detail, the challenge is twofold: (a) identifying IP addresses from other numerical entities, such as software version numbers, and (b) classifying the IP address as benign or malicious. We propose an integrated solution that tackles both these problems. A novelty of our approach is that it does not require training data for each new forum. Our approach does knowledge transfer across forums: we use a classifier from our source forums to identify seed information for training a classifier on the target forum. We evaluate our method using data collected from five security forums with a total of 31K users and 542K posts. First, RIPEx can distinguish IP address from other numeric expressions with 95% precision and above 93% recall on average. Second, RIPEx identifies malicious IP addresses with an average precision of 88% and over 78% recall, using our cross-forum learning. Our work is a first step towards harnessing the wealth of useful information that can be found in security forums.

Keywords: Security, Online communities mining

1 Introduction

The overarching goal of this work is to harness the user generated content in forums, especially security forums. More specifically, we focus here on collecting malicious IP addresses, which are often reported at such forums. We use the term security forums to refer to discussion forums with a focus on security, system administration, and in general systems-related discussions. In these forums, security professionals, hobbyists, and hackers identify issues, discuss solutions, and in general exchange information.

We provide a few examples of the types of discussions that take place in these forums that could involve IP addresses, which is our focus. Posts could talk about a benign IP address, say in configuration files, as in the post: *"[T]his thing in*

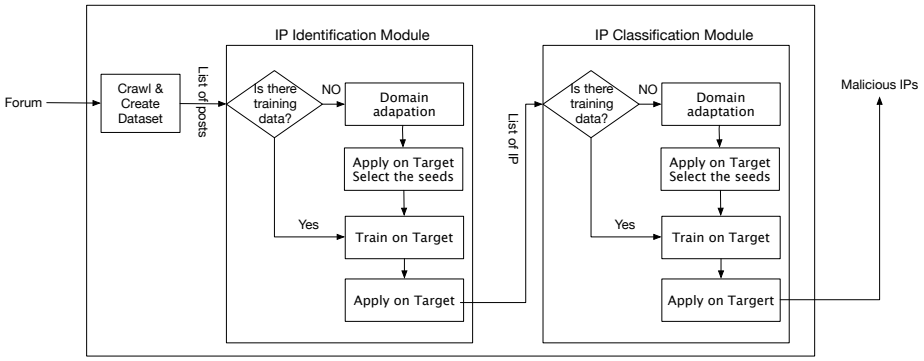


Fig. 1: The overview of key modules of our approach (RIPEx): (a) collecting data, (b) IP Identification, and (c) IP Characterization. In both classification stages, we use our Cross-Seeding approach that in order to generate seed information for training a classifier for a new forum.

my hosts file: 64.91.255.87 ... [is] it correct?". At the same time, posts could also report compromised or malicious IP addresses, as in the post: *"My browser homepage has been hijacked to http://69.50.191.51/2484/"*. Our goal is to automatically distinguish between the two and provide a new source of information for malicious IP addresses directly from the affected individuals.

The problem that we address here is to find all the IP addresses that are being reported as malicious in a forum. In other words, the input is all the posts in a forum and the expected output is a list of malicious IP addresses. As with any classification problem, one would like to achieve both high precision and recall. Precision represents the percentage of the correctly labeled over all addresses labeled malicious. Recall is the percentage of malicious addresses that we find among all malicious addresses reported in forums. It turns out that this is a two-step problem. First, we need to solve the **IP Identification** problem: distinguishing IP addresses from other numerical entities, such as a software version. Second, we need to solve the **IP Characterization** problem: characterizing IP address as malicious or benign. The extent of the Identification problem caught us by surprise: we find 1820 non-address dot-decimals, as we show in table 1.

There is limited work on extracting information from security forums, and even less work on extracting malicious IP addresses. We can group prior work in the following categories. First, recent works study the number of malicious IP addresses in forums, but without providing the comprehensive and systematic solution that we propose here [7]. Second, there are recent efforts that extract other types of information from security forums, related to the black market of hacking services and tools [15], or the behavior and roles of their users [9,1]. Third, other works focus on analyzing structured sources, such as security reports and vulnerability databases [2,11]. We discuss related work in section 5.

There is a wealth of information that can be extracted from security forums, which motivates this research direction. Earlier work suggests that there is close to four times more malicious IP addresses in forums compared to established databases of such IP addresses [8]. At the same time, there are tens of thousands

of IP addresses in the forums, as we will see later. Interestingly, not all of the reported IP addresses are malicious, which makes the classification necessary.

We propose RIPEx[1], a comprehensive, automated solution that can detect malicious IP addresses reported in security forums. As its key novelty, our approach minimizes the need for human intervention. First, once initialized with a small number of security forums, it does not require additional training data to mine new forums. Second, it addresses both the Identification and Characterization problems. Third, our approach is systematic and readily deployable. We are not aware of prior work claiming these three properties, as we discuss in section 5. The overview of our approach is shown in figure 1.

The key technical novelty is that we propose **Cross-Seeding**, a method to conduct a multi-step knowledge transfer across forums. We use this approach for both classification problems, when we have no training data for a new forum. With Cross-Seeding, we create training data for the new forum in the process depicted in figure 1. We use a classifier based on the current forums to identify seed information in the new forum. We then use this seed information to train a classifier for the new forum. This forum-specific classifier performs much better than if we have used the classifier of the current forums on the new forum. We refer to this latter knowledge transfer approach as **Basic**.

We evaluate our approach using five security forums with a total of 31K users and 542K posts spanning a period of roughly six years. Our results can be summarized into the following points.

a. Identification: 98% precision with training data per forum. We develop a supervised learning algorithm for solving the Identification problem in the case where we have training data for the target forum. Our approach exhibits 98% precision and 96% recall on average across all our sites.

b. Identification: 95% precision with Cross-Seeding. We show that our Cross-Seeding approach is effective in transferring the knowledge between forums. Using the WildersSecurity forum as source, we observe an average of 95% precision and 93% recall in the other forums.

c. Characterization: 93% precision with training data per forum. We develop a supervised learning algorithm for solving the Characterization problem assuming we have training data for the target forum. Our classifier achieves 93% precision and 92% recall on average across our forums.

d. Characterization: 88% precision on average with Cross-Seeding data. We show that our Cross-Seeding approach by using OffensiveCommunity forum as source can provide 88% precision and 82% recall on average.

e. Cross-Seeding outperforms Basic. We show that Cross-Seeding is important, as it increases the precision by 28% and recall by 16% on average in the Characterization problem, and the precision by 8% and recall by 7% on average in the Identification problem.

f. Using more source forums improves the Cross-Seeding performance. We show that, by adding a second source forum, we can improve the precision by 13% on average over the remaining three forums.

[1] In the spirit of double-blind reviewing, we withhold the explanation of the acronym.

	WildersSec.	OffensiveComm.	HackThisSite	EthicalHackers	Darkode
Posts	302710	25538	84125	54176	75491
Threads	28661	3542	8504	8745	7563
Users	14836	5549	5904	2970	2400
Dot-decimal	4325	7850	1486	1591	1097
IP found	3891	6734	1231	1330	1082

Table 1: The basic statistics of our forums

Our work suggests that there is a wealth of information that we find in security forums and offers a systematic approach to do so.

2 Our Forums and Datasets

We have collected data from five different forums, which cover a wide spectrum of interests and intended audiences. We present basic statistics of our forums in Table 1 and we highlight the differences of their respective communities.

Our semi-automated crawling tool. We have developed an efficient and customizable python-based crawler, which can be used to crawl online forums, and it could be of independent interest. To crawl a new forum, our tool requires a configuration file that describes the structure of the forum. Leveraging our current configuration files, the task of crawling a new forum is simplified significantly. Due to space limitations, we do not provide further details. Following are the descriptions of collected forums.

- **WildersSecurity (WS)** seems to attract system administrator types and focuses on defensive security: how one can manage and protect one's system. Its topics include anti-virus software, best practices, and new vulnerabilities and its users seem professional and eloquent.
- **OffensiveCommunity (OC)** seems to be on the fringes of legality. As the name suggests, the forum focuses on breaking into systems: it provides step by step instructions, and advertises hacking tools and services.
- **HackThisSite (HT)** seems to be in between these extremes represented by the first two forums. For example, there discussions and competitions on hacking challenges, but it does not act as openly as a black market of illegal services and tools compared to OffensiveCommunity.
- **EthicalHackers (EH)** seems to consist mostly of "white hat" hackers, as its name suggests. The users discuss hacking techniques, but they seem to have a strict moral code.
- **Darkode (DK)** is a forum on the dark web that has been taken down by the FBI in July 2015. The site was a black market for malicious tools and services similar to OffensiveCommunity.

Our goal is to identify and report IP addresses that the forum readers report as malicious. We currently do not assess whether the author of the post is right, though the partial overlap with blacklisted IPs indicates so. We leave for future work to detect misguided reports of IP addresses.

Determining the ground-truth. For both of the problems we address here, there are no well-established benchmarks and labeled datasets. To train and validate our approach, we had to rely on external databases and some manual labelling. For the Identification problem, we could not find any external sources of information and benchmarks. To establish our ground-truth, we selected dot-decimal expressions uniformly randomly, and we used four different individuals for the labelling. To ensure testing fairness, we opted for balanced datasets, which led us to a corpus of 3200 labeled entries across all our forums.

For the Characterization problem, we make use of the VirusTotal site which maintains a database of malicious IP addresses by aggregating information from many other such databases. We also provide a second level of validation via manual inspection.

We create the ground truth by uniformly randomly selecting and assessing IP addresses from our forums. If VirusTotal and the manual inspection give it the same label, we add the addresses into our ground-truth. Finally, we again ensure that we create balanced sets for training and testing to ensure proper training and testing.

3 Overview of RIPEx

We represent the key components of our approach in addressing the Identification and Characterization problems. To avoid repetitions, we present at the end the Cross-Seeding approach, which we use in our solution to both problems.

3.1 The IP Identification module

We describe our proposed method to identify IP addresses in the forum.

The IP address format. The vast majority of IP addresses in the forums follow the *IPv4* dot-decimal format, which consists of 4 decimal numbers in the range [0-255] separated by dots. We can formally represent the dot-decimal notation as follows: *IPv4* $[x_1.x_2.x_3.x_4]$ with $x_i \in [0-225]$, for $i = 1, 2, 3, 4$. Note that the newer *IPv6* addresses consists of eight groups of four hexadecimal digits, and our algorithms could easily extend to this format as well. Interestingly, we found a negligible number of *IPv6* addresses, and we opted to not focus on *IPv6* addresses here. For example, in WildersSecurity forum, we find 3891 *IPv4* addresses and only 56 *IPv6* addresses. At such small numbers, it is difficult to train and test a classifier. Thus, for the rest of this paper, IP address refers to *IPv4* addresses.

The challenge: the dot-decimal format is not enough. If IP addresses were the only numerical expressions in the forums with this format, the Identification problem could have been easily solved with straightforward text processing and Named-Entity Recognition (NER) tools, such as the Stanford NER models [6]. However, there is a non-trivial number of other numerical expressions, which can be misclassified as addresses. For example, we quote a real post: *"factory reset brings me to the Clockworkmod 2.25.100.15 recovery menu"*. where the structure *2.25.100.15* refers to the version of Android app *"Clockworkmod"*.

To this end, we propose a method to solve the IP Identification problem, a supervised learning algorithm. We first identify the features of interest as we discuss below. We then train a classifier using the Logistic Regression method gives the best results among the several methods using 10-fold cross validation on our ground-truth as we decribed in the previous section.

Feature selection. We use three sets of features in our classification.

a. Contextual information: *TextInfo*. Inspired by how a human would determine the answer, we focus on the words surrounding the dot-decimal structure. For example, the words *"server"* or *"address"* suggests that the dot-decimal is an address, while the words *"version"* or a software name, like *"Firefox"* suggests the opposite. At the same time, we wanted to focus on words close to the dot-decimal structure. Therefore, we introduce **Word-Range,** W**,** to determine the number of surrounding words before and after the dot-decimal structure that we want to consider in our classification. We use TF-IDF [16] to normalize the frequency of a word to better estimate its discriminatory value.

b. The numerical values of the dot-decimal: *DecimalVal*. We use the numerical value of the four numbers in the the dot-decimal structure as features. The rationale is that non-addresses, such as software versions, tend to have lower numerical values. This insight was based on our close interaction with the data.

c. The combined set: *Mixed*. We combine the two feature sets to create in order to leverage their discriminating power.

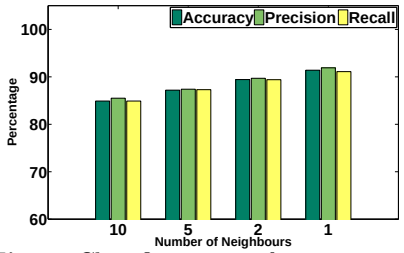


Fig. 2: Classification performance versus the number of words Word-Range, W, in WildersSecurity.

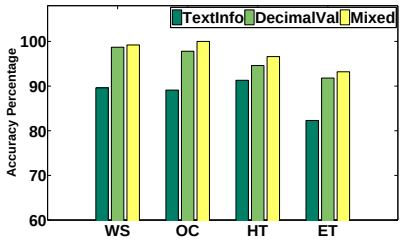


Fig. 3: Classification accuracy for different features sets in 10-fold cross validation in four forums.

Determining the right number of context words, Word-Range. We wanted to identify the best value of parameter Word-Range for our classification. In figure 2, we plot the classification accuracy, precision and recall, as we vary Word-Range, $W = 1, 2, 5$ *and* 10, for the WildersSecurity forum and using only the *TextInfo*. We see that using one to two words gives better results compared to using five and ten words. The explanation to this counter-intuitive result is that considering more words includes text that is not relevant for inferring the nature of a dot-decimal, which we verified manually.

Using numerical values *DecimalVal* improves the performance significantly. In Figure 3, we plot the classification accuracy of different features sets. Recall that we are not able to include Darkode forum due to its limited number of non-IP dot-decimal expressions, as we saw in 3.1. We see that us-

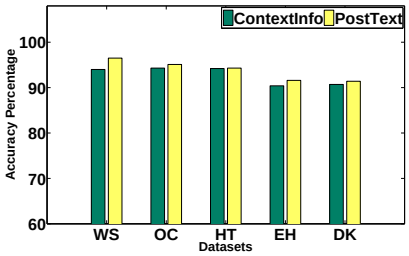


Fig. 4: Characterization: The effect of the features set on the classification accuracy with balanced testing data.

ing *DecimalVal* features alone, we can get 94% overall accuracy and using both *DecimalVal* and *TextInfo*, we get 98% overall accuracy across our forums. Focusing on the IP address class, we see a an average precision of 95% using only *DecimalVal* and, 98% using both *DecimalVal* and *TextInfo*.

3.2 The IP Characterization module

We develop a supervised learning algorithm to characterize IP addresses. Here, we assume that we have labeled data, and we discuss how we handle the absence of ground truth in section 3.3. We first identify the appropriate set of features which we discuss below. We then train a classifier and find that the Logistic Regression method gives the best results among several methods that we evaluated. Due to space limitations, we show a subset of our results.

Features sets for the Characterization problem. We consider and evaluate three sets of features in our classification.

a. Text information of the post: *PostText*. We use the words and their frequency of appearance in the post. Here, we use the TF-IDF technique [16] again to better estimate the discriminatory value of a word by considering its overall frequency. In the future, we intend to experiment with sophisticated Natural Language Processing models for analyzing the intent of a post.

b. The Contextual Information set: *ContextInfo*. We consider an extended feature set that includes both the *PostText* features, but also features of the author of the post. These features capture the behaviour of the author, including frequency of posting, average post length etc. These features were introduced by earlier work [8], with the rationale that profiling the author of a post can help us infer their intention and role and thus, improve the classification.

Characterization: 93% precision with training data. We assess the performance of the Characterization classifier using the set of features above and by using the labeled data of each forum. We evaluate the performance using 10-fold cross validation. In figure 4, we show the accuracy of classification.

We can achieve 93% precision and 92% recall on average across all the forums. The results are shown in figure 4, where we report the results using the accuracy across both classes, given that we have balanced training datasets.

Selecting the *PostText* feature set. We see that, by using *PostText* features on their own, we obtain slightly better results. *PostText* feature achieves 94% accuracy on average, while using the *ContextInfo* results in 92% accuracy

on average across all forums. Furthermore, text-based only features have one more key advantage: they can transfer between domains in a straightforward way. Therefore, we use the *PostText* features in the rest of the paper.

3.3 Transfer Learning with Cross-Seeding

In both classification problems, we face the following conundrum:

a. the classification efficiency is better when the classifier is trained with forum-specific ground-truth, but,

b. requiring ground-truth for a new forum will introduce manual intervention, which will limit the practical value of the approach.

We propose to do cross-forum learning by leveraging transfer learning approaches [5,14]. We use the terms *source* and *target* domain to indicate the two forums with the target forum not having ground-truth available. For both classification problems, we consider two solutions for classifying the target forum:

a. Basic: We use the classifier from the source forum on the target forum.

b. Cross-Seeding: We propose an algorithm that will help us develop a new classifier for the target forum by using the old classifier to create training data as we explain below.

Algorithm 1: Cross-Seeding: transfer learning between forums

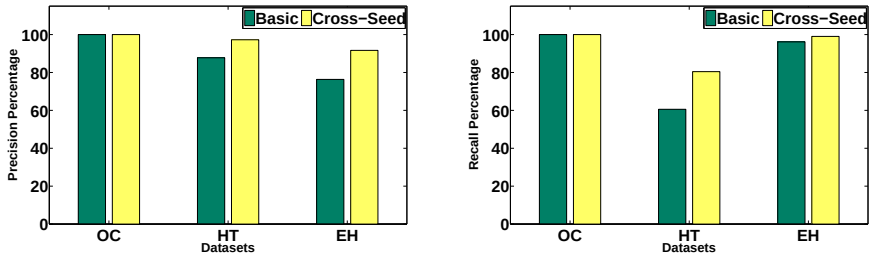
```
1 CrossForum (𝒳, 𝒴) :
2     Take the union of the features in forum 𝒳 and 𝒴
3     Apply classifier from 𝒳 on 𝒴
4     Select the high-confidence instances to create seed for 𝒴
5     Train a new classifier on 𝒴 based on the new seeds.
6     Apply the new classifier on 𝒴
```

Our Cross-Seeding approach. We propose to create training data for the target forum following the four steps below, which are illustrated in figure 1 and outlined in algorithm 1.

a. Domain adaptation. The main role of this step is to ensure that the source classifier can be applied to the target forum. The main issue in our case is that the feature sets can vary among forums. Recall that, for both classification problems, we use the frequency of words and these words can vary among forums. We adopt an established approach that works well for text classification [5]: we take the union of the feature sets of the source and target forums. The approach seems to work sufficiently well in our case, as we see later.

b. Creating seed information for the target forum. Having resolved any potential feature disparities, we can now apply the classifier from the source forum to the target forum. We create the seeding data by selecting instances of the target domain, for which the classification confidence is high. Most classification methods provide a measure of confidence for each classified instance and we revisit this issue in section 4.

c. Training a new classifier for the target forum. Having the seed information, this is now a straightforward step of training a classifier.



(a)Precision (b)Recall

Fig. 5: Identification: Cross-Seeding improves both Precision and Recall. Using WildersSecurity to classify OffensiveCommunity, HackThisSite, and EthicalHackers.

d. Applying the new classifier on the target forum. In this final step, we apply our newly-trained forum-specific classifier on the target forum.

4 Evaluation of our Approach

We evaluate our approach focusing on the performance of Cross-Seeding for both the Identification and the Characterization problems.

Our classifier. We use Logistic Regression as our classification engine, which performed better than several others, including SVM, Bayesian networks, and K-nearest-neighbors. In Cross-Seeding, we use the Logistic Regression's prediction probability with a threshold of 0.85 to strike a balance between sufficient confidence level and adequate number of instances above that threshold. We found this value to provide better performance than 0.8 and 0.9, which we also considered.

A. The IP Identification problem. As we saw in section 3.1, our classification approach exhibits 98% precision and 96% recall on average across all our sites, when we train with ground-truth for each forum.

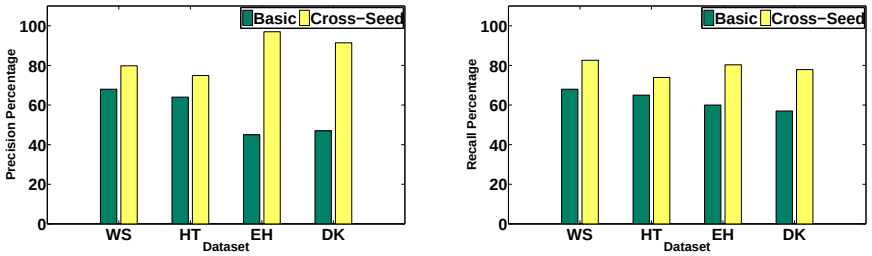
a. Identification: 95% precision with Cross-Seeding. We show that our cross-training approach is effective in transferring the knowledge between domains. We use the classifier from WildersSecurity and we use it to classify three of the other forums, namely, OffensiveCommunity, EthicalHackers, and HackThisSite. Note that we do not include Darkode in this part of the evaluation as it did not have sufficient data for testing (less than 15 non-address expressions in all its posts).

In figure 5, we show the results for precision and recall of cross-training using Basic and Cross-Seeding. We see that Cross-Seeding improves *both* precision and recall significantly. For example, for HackThisSite, Cross-Seeding increases the precision from 57% to 79% and the recall from 60% to 78%.

b. Identification: Cross-Seeding outperforms Basic. Cross-Seeding improves the precision by 8% and recall by 7% on average for the experiment shown in figure 5. The average precision increased from 88% to 95% and the average recall increased from 85% to 97%.

B. The IP Characterization problem. We evaluate our approach for solving the Characterization problem without per-forum training data. As we

saw in section 3.2, we can achieve 93% precision and 92% recall on average across all the forums, when we train with ground-truth for each forum.



(a)Precision (b)Recall

Fig. 6: Characterization: Cross-Seeding improves both Precision and Recall. Using OffensiveCommunity as source, we classify WildersSecurity, HackThisSite, EthicalHackers and Darkode.

a. Characterization: 88% precision on average with Cross-Seeding. Using OffensiveCommunity as source, and we classify WildersSecurity, HackThisSite, EthicalHackers and Darkode as shown in figure 6. Our Cross-Seeding approach can provide 88% precision and 82% recall on average.

b. Characterization: Cross-Seeding outperforms Basic. We show that Cross-Seeding improves the classification compared to just reusing the classifier from another forum. In figure 6, we show the precision and recall of the two approaches. Using OffensiveCommunity as our source, we see that Cross-Seeding improves the precision by 28% and recall by 16% on average across the forums compare to the Basic approach. We also observe that the improvement is substantial: Cross-Seeding improves both precision and recall in all cases.

	OffensiveComm.	HackThisSite	Darkode	Average
Precision	3.3	20.5	17.8	13.2
Recall	8.3	6.4	38.8	17.8

Table 2: Characterization: Using two instead of one source forums improves precision and recall on average: Average improvement of using EthicalHackers and WildersSecurity as sources together compared to each of them individually.

c. Using more source forums improves the Cross-Seeding performance significantly. We quantify the effect of having more than one source forums in the classification accuracy of a new forum. We use EthicalHackers and WildersSecurity as our training forums, and we use Cross-Seeding for OffensiveCommunity, HackThisSite, and Darkode. First, we use the source forums one at a time and then both of them together. In table 2, we show the average improvement of having two source forums over having one for each target website. Using two source forums increases the classification precision by 13% and the recall by 17% on average.

Discussion: Source forums and training. How would we handle a new forum? Given the above observations, we would currently use all our five forums as sources for a new forum. Overall, we can argue that the more forums we have,

the more we can improve our accuracy. However, we would like to point out that some forums are more “similar" and thus more suitable for cross-training. We will investigate how to best leverage a large group of source forums once we collect 20-25 more forums.

5 Related Work

We summarize related work clustered into areas of relevance.

a. Extracting IP addresses from security forums. There two main efforts that focus on IP addresses and security forums [7,8] and neither provides the comprehensive solution that we propose here. The most relevant work [8] does not address the Identification problem, and sidesteps the problem of cross-forum training by assuming training data for each forum. The earlier work [7] focuses on the spatiotemporal properties of Canadian IP addresses in forums, but assumes that all identified addresses are suspicious and therefore they did not employ a classification method, which is the focus of our work.

b. Extracting other information from security forums. Various efforts have attempted to extract other types of information from security forums. A few recent studies identify malicious services and products in security forums by focusing on their availability and price [15,13].

c. Studying the users and posts in security forums. Other efforts study the users of security forums, group them into different classes, and identify their roles and social interactions [1,20,9,17,18].

d. Analyzing structured security-related sources. There are several studies that automate the extraction of information from structured security documents, extracting ontology and comparing the reported information, such as databases of vulnerabilities, and security reports from the industry [12,2,10].

Transfer learning methods and applications. There is extensive literature on transfer learning [5,4,3] and several good surveys [14,19], which inspired our approach. However, to the best of our knowledge, we have not found any work that address the same domain-specific challenges or uses all the steps of our approach, which we described in 3.3.

6 Conclusion

We propose a comprehensive solution for mining malicious IP addresses from security forums. A novelty of our approach is it minimizes the need for human intervention. First, once it is initialized with a small number of security forums, it does not require additional training data for each new forum. To achieve this, we use Cross-Seeding, which uses initialization via domain adaptation: we use a classifier from current forums to create seed information for the new forum. Second, it addresses both the Identification and Characterization problems, unlike all prior work that we are aware of. We evaluate our method real data and we show that: (a) our Cross-Seeding approach works fairly well reaching precision above 85% on average for both classification problems, (b) Cross-Seeding outperforms the Basic approach, and (c) using more source forums increases the performance as one would expect.

Our future plans include: (a) collecting a large number of security forums, (b) exploring the limits of the classification accuracy by using more source forums, and (c) exploring additional transfer learning methods.

References

1. A. Abbasi, W. Li, V. Benjamin, S. Hu, and H. Chen. Descriptive analytics: Examining expert hackers in web forums. ISI '14, pages 56–63, 2014.
2. R. A. Bridges, C. L. Jones, M. D. Iannacone, and J. R. Goodall. Automatic labeling for entity extraction in cyber security. *CoRR*, abs/1308.4941, 2013.
3. W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Co-clustering based classification for out-of-domain documents. KDD '07, pages 210–219, USA, 2007.
4. W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. ICML '07, pages 193–200, New York, NY, USA, 2007.
5. H. Daume III. Frustratingly easy domain adaptation. ACL '07, 2007.
6. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. ACL '05, 2005.
7. R. Frank, M. Macdonald, and B. Monk. Location, location, location: Mapping potential canadian targets in online hacker discussion forums. EISIC '16, 2016.
8. J. Gharibshah, T. C. Li, M. Solanas Vanrell, A. Castro, K. Pelechrinis, E. Papalexakis, and M. Faloutsos. Inferip: Extracting actionable information from security discussion forums. ASONAM '17, 2017.
9. T. J. Holt, D. Strumsky, O. Smirnova, and M. Kilger. Examining the social networks of malware writers and hackers. 6(1):891–903, 2012.
10. M. Iannacone, S. Bohn, G. Nakamura, J. Gerth, K. Huffer, R. Bridges, E. Ferragut, and J. Goodall. Developing an ontology for cyber security knowledge graphs. CISR '15, pages 12:1–12:4, New York, NY, USA, 2015.
11. C. L. Jones, R. A. Bridges, K. M. T. Huffer, and J. R. Goodall. Towards a relation extraction framework for cyber-security concepts. CISR '15, 2015.
12. C. L. Jones, R. A. Bridges, K. M. T. Huffer, and J. R. Goodall. Towards a relation extraction framework for cyber-security concepts. *CoRR*, abs/1504.04317, 2015.
13. M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. An analysis of underground forums. IMC '11, pages 71–80, New York, NY, USA, 2011.
14. S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
15. R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson. Tools for automated analysis of cybercriminal markets. WWW '17, 2017.
16. J. Ramos. Using TF-IDF to determine word relevance in document queries. ICML '03, 2003.
17. S. Samtani, R. Chinn, and H. Chen. Exploring hacker assets in underground forums. ISI '15, pages 31–36, 2015.
18. J. Shakarian, A. T. Gunn, and P. Shakarian. *Exploring Malicious Hacker Forums*, pages 259–282. Springer International Publishing, Cham, 2016.
19. K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016.
20. X. Zhang, A. Tsang, W. T. Yue, and M. Chau. The classification of hackers by knowledge exchange behaviors. *Info. Systems Frontiers*, 17(6), 2015.